Chromosome-level de novo genome assembly of Sarcophaga peregrina provides insights into the evolutionary adaptation of flesh flies

Lipin Ren[1], Yanjie Shang[2], Li Yang[2], Shiwen Wang[3], Xiang Wang[4], Shan Chen[5], Zhigui Bao[6], Dong An[6], Fanming Meng[1], Jifeng Cai[2], and Yadong Guo[2]

[1]Affiliation not available
[2]Central South University
[3]Xin Jiang Medical University
[4]Dalian Medical University
[5]East China Normal University
[6]OE biotech Co Ltd

May 5, 2020

Abstract

Sarcophaga peregrina is usually considered to be of great ecological, medical and forensic significance, and has the biological characteristics such as the ovoviviparous reproductive pattern and adaptation to feed on carrion. However, the underlying mechanisms still remain unsolved by lack of high-quality genome. Here we present de novo–assembled genome at chromosome-scale for S. peregrina. The final assembled genome was 560.31 Mb with contig N50 of 3.84 Mb. Hi-C scaffolding reliably anchored six pseudochromosomes, accounting for 97.76% of the assembled genome. Moreover, 45.70% of repeat elements were identified in the genome. A total of 14,476 protein-coding genes were functionally annotated, accounting for 92.14% of all predicted genes. Phylogenetic analysis indicated that S. peregrina and S. bullata diverged ~7.14 Mya. Comparative genomic analysis revealed expanded and positively selected genes related to biological features that aid in clarifying its ovoviviparous reproduction and necrophagous habit, such as horionic membrane formation and Dorso-ventral axis formation, lipid metabolism, and olfactory receptor activity. This study provides a valuable genomic resource of S. peregrina, and sheds insight into further revealing the underlying molecular mechanisms of adaptive evolution.

KEYWORDS

Sarcophaga peregrina , *de novo* genome assembly, comparative genomics, adaptive evolution

1 INTRODUCTION

Sarcophaga peregrina (Robineau-Desvoidy, 1830) (Diptera: Sarcophagidae), commonly known as flesh fly, which is closely associated with human life in ecological habits. The species is widely spread from tropical to subtropical areas of the Palaearctic, Oriental, and Oceanian regions (Xue *et al.* 2011). Moreover, it is also a large-sized flesh fly with significant body surface features, including the brightly red-tipped eyes, gray and black longitudinal stripes on the thorax, and a checkerboard-like pattern on the abdomen (Majumder *et al.* 2012) (Additional file 1: Figures S1).

The common carrion-feeding flies mainly include Sarcophagidae, Calliphoridae and Muscidae family, which play a crucial role in forensic investigations associated with decomposed corpses (Byrd& Castner 2010). Compared with the other necrophagous flies, flesh flies are characterized by the reproductive pattern of ovoviviparity (or ovolarviparity), depositing eggs which immediately hatch into larvae onto carrion (Goff *et al.* 1989; Majumder *et al.* 2014). The reproductive cycle of *S. peregrina* comprises three definite stages, including larva, pupa and adult. The mode of reproduction appears to be the result of adaptive evolution, making them more competitive compared to other species. Given that this reproduction reduces the stage of larval development (the time when eggs hatch to first larvae), the species can be used very accurately to estimate the postmortem interval (PMI) of decomposed corpses, and it is therefore an important necrophagous flesh fly in the field of forensic entomology (Byrd& Castner 2010). For instance, *S. peregrina* is one of the most common species of insect succession patterns on cadavers as well as at many death scenes colonizing on a corpse (Guo *et al.* 2014; Siti Aisyah *et al.* 2015; Sukontason *et al.* 2010; Wang*et al.* 2017a), which would provide valuable data for forensic investigations, especially floating corpse cases and indoor death-scene (Tomberlin *et al.* 2011). Recent studies on the species have mainly focused on the effect of drugs and heavy metals (eg. cadmium) on growth and development of larvae (Goff *et al.* 1989; Wu *et al.* 2013), molecular identification (Wells& Stevens 2008), larval morphology (Sukontason *et al.* 2010), cuticular hydrocarbon composition in pupal exuviae for taxonomic differentiation (Gongyin*et al.* 2007), as well as the developmental data collection at constant temperatures (Wang *et al.* 2017b).

S. peregrina has also profound implications for human hygiene and the livestock economy. It is an important sanitary insect pest and one of the vector fly species of intestinal infectious diseases and parasitic diseases in human and livestock, and as an ectoparasite causing myiasis (parasitic infestation) in human and other mammals (Lee*et al.* 2011). They can cause myiasis in the hospital environment which is also called nosocomial myiasis (Miura *et al.* 2005). As such, the species is considered as an indicator of wound care neglect, either by the nurses or oneself (Nazni *et al.* 2011). Additionally, in off-shore islands, the larvae of this species are also the key pest of meat industries, which take nutrient from uncovered meat and contaminate food material, ultimately leading to economic losses in the livestock industries (Majumder *et al.* 2012).

Although *S. peregrina* possesses ecological, medical and forensic importance, there are few genomic resources for the family Sarcophagidae (Agrawal *et al.* 2010; Martinson *et al.* 2019), which seriously hinder the investigation of the specific mechanisms of biological phenomena from the perspective of genomics, transcriptomics and epigenetics. Fortunately, with the emergence of next-generation sequencing technology, genomics and transcriptomic have been recently developed for dipteran flies (Anstead *et al.* 2015; dos Santos*et al.* 2014; Kim *et al.* 2018; Scott *et al.* 2014), which serve as a reference for molecular studies of related species. But due to the defect of short Illumina reads, the combination of SMRT (Single Molecule Real-Time) sequencing and chromosome conformation capture (Hi-C) can anchor the scaffolds into chromosomal levels (Belton *et al.* 2012; Roberts *et al.* 2013), which ensure the availability of high-quality reference genome assembly. Here we reported a chromosome-level *de novo* genome assembly of *S. peregrina*and perform comparative analysis with other published dipteran insects in order to enrich our understanding of adaptive evolution in *S. peregrina* .

2 MATERIALS AND METHODS

2.1 Cultivation of *S. peregrina*

Adult specimens of *S. peregrina* were trapped with pork liver bait in Changsha, Hunan Province, China, and raised at 25 ± 1°C and 70 ± 5% relative humidity with a photoperiod regime of 12:12h light/darkness in an artificial climate chamber, employing pork liver as a medium for larvipositing and larval rearing. In order to reduce genetic variability, mating pairs of adult *S. peregrina* trapped in the wild were highly inbred for six generations. Afterwards, 3rd-instar larvae and newly hatched adult females were used for further studies.

2.2 Genome survey

Genomic DNA was extracted from a single adult female of *S. peregrina* using SDS method (Rinkevich *et al.* 2006). A library with insert sizes of 400 bp was generated by Illumina TruSeq Nano DNA Library Prep Kit and sequenced to 150-nt paired end reads (PE 150 bp) on the Illumina HiSeq Xten (San Diego, CA, USA). After quality control, low-quality reads, sequencing adapters, contaminated reads and ambiguous bases were removed, whilst duplicates were filtered out. Finally, the clean data were applied to enable the genome survey and calibrate subsequent genomic assembly. Meanwhile, in order to estimate the genome size and heterozygosity of *S. peregrina* , we performed the$K$ -mer distribution ($K = 17$) using the Jellyfish program (Marcais& Kingsford 2011).

2.3 Genome sequencing

Genomic DNA was extracted from a single adult female of *S. peregrina* using SDS method (Additional file 1: Table S1) (Rinkevich*et al.* 2006). A 20-kb insert SMRTbell library was generated using the SMRT bell Template Prep Kits according to the selection protocol on the BluePippin (Sage Science, MA, USA). SMRT sequencing was performed on PacBio Sequel instrument with 11 SMRT cells at the Genome Center of Nextomics (Wuhan, China). After removal of low quality and sequencing adapters, the clean subreads were used for subsequent genome assembly. Furthermore, in order to assist genome annotation, total RNAs were extracted from a whole single adult female using the mirVana miRNA Isolation Kit (Ambion) following the manufacturer's protocol. RNA-seq libraries were then constructed using TruSeq RNA Library Preparation Kit (Illumina, CA, USA), and paired-end sequencing (PE 150 bp) was performed on HiSeq Xten platform. A total of 9.62 Gb of raw data was produced for assisting genome annotation (Additional file 1: Table S2).

2.4 *De novo* assembly and polishing of the genome

PacBio long subreads were originally corrected with Canu v1.6 (Koren*et al.* 2017). The genome assembly was performed on WTDBG v1.2.8 using the error-corrected reads. The PacBio Subreads were subsequently mapped back to the raw contigs by Blasr v5.1, and contigs were further polished in Arrow v2.1.0 (Chin *et al.* 2013). Due to a high error ratio of PacBio raw long reads, Illumina short reads were mapped back to the improved contigs and further polished by Pilon v1.20 (Walker*et al.* 2014). In addition, we applied the GC depth analysis to evaluate whether potential contamination remained during sequencing and the coverage of the assembly. The analysis showed that an average GC content of the genome was 33.23% and a single-peaked distribution cure (Additional file 1: Figures S2 and S3). Combining the GC depth analysis with the sequencing depth of the genome indicated that there was no contamination from other species (Additional file 1: Figure S4).

2.5 Hi-C library preparation

The Hi-C library preparation was constructed following a previous protocol (Zhuang *et al.* 2019), Briefly, A single 3rd-instar larva was washed, and after removal of the dissected guts, the remaining tissues were fixed with 1% formaldehyde. Subsequently, glycine was added to quench the cross-linking reaction, and tissues were finally ground to powder and suspended in nuclei isolation buffer to obtain a nuclei suspension (Belton *et al.* 2012). The nuclei mixture was dissolved, the chromatin was digested with restriction enzyme (Dpn II), and end-labeled by incubating with Klenow enzyme and biotin-14-dCTP generating blunt-end-repaired DNA strands (Lieberman-Aiden *et al.*2009), and ligated by T4 DNA polymerase. The extracted DNA was mechanically sheared to 200–300 bp sizes. DNA fragments of 150–300 bp were blunt-end repaired and A-tailed, followed by purification through biotin–streptavidin-mediated pulldown (Burton *et al.* 2013). PCR amplification was performed after adapters were ligated to the Hi-C products. The PCR products were purified with AMPure XP beads, and the Hi-C libraries were quantified by quantitative PCR reaction (Lieberman-Aiden *et al.* 2009). The Hi- C library was constructed by the NEBNext Ultra II DNA library Prep Kit and then sequenced (150 bp paired-end reads) on the Illumina HiSeq Xten.

2.6 Pseudomolecule construction by Hi-C

To construct a chromosomal-level assembly of the genome, Hi-C raw data were first trimmed by fastp v. 0.12.6 (Chen *et al.* 2018). After quality control, the low-quality reads, adapter contamination and ambiguous bases as N's were removed, whilst duplicates were filtered out. High quality clean paired-end reads were retained. The clean paired-end reads were aligned with the draft assembled genome using Juicer pipeline v. 2.3.2 (Durand *et al.* 2016). Afterwards, according to the location of DpnII restriction sites, the ratio of Self Circle, Dangling End and Dumped Pairs was identified so as to evaluate the validity of the paired-end reads. The contigs were then clustered, ordered and oriented using the 3D de novo assembly (3d-DNA,

v. 170 123) pipeline (Dudchenko *et al.* 2017). Hi-C contact matrix was visualized using Juicebox v. 1.9.8 (Dudchenko *et al.* 2018; Robinson *et al.* 2018). The misassembly and misconnection were manually adjusted based on neighboring interactions. The validated assembly was used to construct pseudomolecules using the finalize-output.sh script from 3d-DNA. Meanwhile, the completeness of the assembly was evaluated using Benchmarking Universal Single-Copy Orthologs (BUSCO) v3.1 (Simão *et al.* 2015).

2.7 Gene prediction and functional annotation

In order to identify the tandem repeats and transposable elements (TEs) in repetitive sequences, we combined the *de novo* and homolog-based methods. A *de novo* specific repeat library was first generated using the RepeatModeler v. 1.0.11 (Bedell *et al.*2000). The repetitive sequences in the assembled genome were annotated by the RepeatMasker v 4.0.6 with default parameters (Bedell *et al.* 2000). Afterwards, RepeatMasker, RepeatProteinMask and Tandem repeats finder (TRF, v 4.09) were used to search against the known RepBase repeats (Allred *et al.* 2008; Bedell *et al.* 2000; Benson 1999; Jurka *et al.*2005). In addition, the simple sequence repeats (SSRs) were identified as implemented in MIcroSAtellite Identification Tool (Thiel *et al.* 2003).

The non-coding RNAs (ncRNAs) were annotated using BLAST (E-value [?] 1e-5) from the Rfam database (Camacho *et al.* 2009; Kalvari*et al.* 2018), including microRNAs (miRNAs), ribosomal RNAs (rRNAs), snRNAs and transfer RNAs (tRNAs). RNAmmer v1.2 was used to predict the rRNAs and their subunits (Lagesen *et al.* 2007). We also annotated the tRNAs by tRNAscan-SE v1.3.1 with default parameters (Lowe& Eddy 1997).

We combined homology searches, *de novo* prediction and transcriptome data-based approaches to predict protein-coding gene structures of *S. peregrina* . In the homology-based method, protein sequences from five dipteran insects (*Aedes aegypti* ,*Anopheles gambiae* , *Drosophila melanogaster* , *Lucilia cuprina* , and *Musca domestica*) were used as queries to search against the assembled genome using the GeneWise v2.4.1 (Birney& Durbin 2000). The *de novo* predictions were performed from the homology-based predictions to train model parameters using the Augustus v3.0 (Stanke *et al.* 2004), SNAP (Korf 2004), GlimmerHMM (Majoros*et al.* 2004), and GeneID v1.4.4 (Bromberg& Rost 2007). Meanwhile, transcriptome data was utilized to align against the genome assembly through PASA and TopHat, respectively (Haas *et al.* 2008; Moriya *et al.* 2007). Subsequently, we integrated all predicted genes to generate a consensus gene set via EVidenceModeler v1.1.1 (Haas*et al.* 2008). The genes containing TEs were then abandoned using the TransposonPSI package to search against the Repbase (Yagi *et al.* 2013). Finally, all gene sets were predicted in assembled genome. Additionally, in order to annotate gene functions, the predicted genes were aligned against the NR, Swissprot, TrEMBL, KEGG, KOG, GO, Pfam and InterProscan databases.

2.8 Gene family analysis

In order to identify gene families, protein sequences from *S. peregrina* and other nine diptera species, including *Sarcophaga bullata* , *Stomoxys calcitrans* , *M. domestica* , *L. cuprina* , *Ceratitis capitata* , *Bactrocera oleae* and*D. melanogaster* , with *A. aegypti* and *A. gambiae*being the family Culicidae as an outgroup, were first aligned using DIAMOND v0.9.30 (Buchfink *et al.* 2015), and the aligned results were then clustered using OrthoFinder v2.7 (Emms& Kelly 2015; Xu*et al.* 2019). To further reveal the phylogenetic relationships among *S. peregrina* and other nine species mentioned above, single-copy families were aligned via the MAFFT v7.0 (Kazutaka& Standley 2013), and then trimmed using the Gblocks v0.91b (Castresana 2000). The phylogenetic tree was inferred using a maximum likelihood method as implemented in RAxML v8.2 with the GTRGAMMA model and 100 bootstrap replicates (Alexandros 2006; Stamatakis 2014). Afterwards, divergence times were estimated under a relaxed clock model by MCMCTREE program implemented in PAML v 4.9e (Yang 2007). The molecular clocks of the family Culicidae (105.91-234.53 Mya), *Stomoxys calcitrans* and*M. domestica* (26.97-36.96 Mya) were used for fossil calibration.

According to gene families and phylogenetic relationships, the results were further analyzed to identify the expanded and contracted gene families by Computational Analysis of Gene Family Evolution v4.2 (Tijl*et al.* 2006). Moreover, in order to identify positively selected genes in the *S. peregrina* , we retained orthologous groups among*S. peregrina* and the remaining seven species (after removal of the outgroup in evolutionary analysis) using Blastall (Camacho *et al.* 2009). Subsequently, we calculated likelihood ratio tests for selection (P<0.05) using Codeml with the branch-site model as implemented in the PAML package (Yang 2007). Besides, we conducted the chromosome synteny between *S. peregrina* and *D. melanogaster* based on genome-scale ortholog alignment using MCScanX (Wang *et al.* 2012).

3 RESULTS

3.1 Genome sequencing and assembly

In order to survey the genome of *S. peregrina* , 46.8 Gb of raw illumina data were produced, of which 40.4 Gb clean data were retained. A total count of *17* -mer was 28,804,585,532 from short clean reads. The distribution curve of *17* -mer presented an unusual Poisson distribution with two upward convex signals, suggesting high heterozygosity. Given that high heterozygosity is also common in other insects, the second peak was selected to be the main peak. According to the peak *17* -mer depth of 61, the genome size was estimated to be ~472 Mb, which was highly heterozygous (~3.0%) (Additional file 1: Table S3 and Figure S5) (Vurture *et al.* 2017). Moreover, we performed the heterozygosity analysis using SNP calling implemented in GATK v. 4.1.5.0 (Walker*et al.* 2018). The heterozygosity ratio was ~1.65%, which is relatively lower than that of*17* -mer analysis (Additional file 1: Table S4).

We generated 58.54 Gb of raw PacBio data. After quality control, 57.83 Gb of subreads were retained for genome assembly. The average length and the N50 of subreads were 8.55 kb and 13.90 kb, respectively (Additional file 1: Table S5). The initial genome assembly was 554.66 Mb in size, with contig N50 of 3.79 Mb and contig number of 2,031, respectively (Additional file 1: Table S6). Finally, the *de novo* genome assembly was 560.31 Mb in size, with contig N50, the longest contig and contig number of 3.84 Mb, 20.90 Mb and 2,031, respectively (Table 1, Additional file 1: Table S6). Meanwhile, the result of completeness of the assembly indicated that the genome assembly covered 97.9% complete BUSCOs and 97.1% of single-copy BUSCOs, with only 1.4% of missing BUSCOs (Additional file 1: Table S7).

A total of 159.4 Gb of Hi-C raw data were produced consisting of 1,063,074,766 paired-end reads (Additional file 1: Table S8), After quality control, 153.8 Gb of clean data were obtained, containing 96.45% of clean paired-end reads (Additional file 1: Table S9), which were used as input for the Juicer and 3d-DNA Hi-C analysis and scaffolding pipelines. Finally, pseudochromosomes with a total length of 548.19 Mb were exactly anchored into six chromosomes, accounting for 97.76% of the draft assembled genome **(Fig. 1)** , which is identical to the karyotype of six chromosomes based on cytological observation in *S. peregrina* (Agrawal *et al.* 2010) (**Fig. 2a** , Additional file 1: Table S10). Although the size of the assembled genome is more than twice that of *D. melanogaster* , six pseudochromosomes in the assembled genome can be aligned nearly against the *D. melanogaster* genome (**Fig. 2b**). The result of completeness of the assembly indicated that the Hi-C genome assembly covered 98.2% complete BUSCOs and 97.4% of single-copy BUSCOs, with only 0.8% of duplicated BUSCOs (Additional file 1: Table S11).

3.2 Genome annotation

A total of 15,710 genes were identified with an average number of 3.94 exons per gene, average transcript length, average CDS length, and average exon length per gene were 7,635.25 bp, 1,404.03 bp, and 356.4 bp, respectively (Additional file 1: Table S12). Moreover, the total number of genes in assembled genome is larger than those of five published genomes mentioned above (*A. aegypti* , *A.s gambiae* , *D. melanogaster* , *L. cuprina* , and *M. domestica*) (Additional file 1: Figure S6 and Table S13). 14,476 protein-coding genes were annotated with potential functions, accounting for 92.14% of all genes in assembled genome (Additional file 1: Table S14). We identified 11,425 genes that showed homology to proteins in the InterPro databases. A total of 7,999 genes were assigned to GO classifications. Based on KEGG analysis, we could annotate 5,236 genes and 130 KEGG metabolic pathways in the assembled genome. Additionally, 9,332 genes could be annotated in the Swissprot database.

The results of the *de novo* and homology-based predictions showed that 256.07 Mb of repetitive sequences were identified, covering 45.70% of the assembled genome. DNA transposons (69.21Mb) represented the most abundant TEs, accounting for 12.35% of the genome (Additional file 1: Table S15 and S16). Furthermore, 456,324 SSRs were detected, including 338,634, 58,559, 43,763, 12,124, 2,835, 409, mono-, di-, tri-, tetra-, penta-, and hexa-nucleotide repeats, respectively (Additional file 1: Tables S17 and S18). In addition, 157 miRNAs, 50 rRNAs, 200 snRNAs, and 1,465 tRNAs were identified in the assembled genome (Additional file 1: Table S19).

3.3 Gene family identification and phylogeny analysis

Finally, 9,636 gene families were identified in assembled genome, covering 13,039 genes. Among these, thirteen gene families containing 106 genes were unique to *S. peregrina* . Besides, 2,662 unclustered genes were identified (**Fig. 2c** , Additional file 1: Table S20). We then identified 5,622 single-copy orthologs to construct phylogenetic trees (Additional file 1: Figure S7 and Table S21). Phylogenetic analysis indicated that eight fly species were clustered together into a large branch and strongly supported (ML bootstrap percentage, BP = 100), whilst *S. peregrina* and *S. bullata* were clustered more closely than other species. As an outgroup taxon, *A. aegypti* and *A. gambiae* (Diptera: Culicidae) are clustered together and clearly separated (Additional file 1: Figure S8). Moreover, the family Sarcophagidae is more closely related to Calliphoridae than to other family, and estimated divergence time between them was 32.53 Mya (95% HPD: 25.38–40.06 Mya) within the Late Paleogene epoch. This is likely consistent with that both of them constitute the main part of insect faunal succession on decomposed remains (Byrd& Castner 2010). Moreover, within the family Sarcophagidae, the diversification of *S. peregrina* and *S. bullata* took place 7.14 Mya (95% HPD: 4.99–9.43 Mya) (Additional file 1: Figure S9).

3.4 Gene family expansion and contraction

This approach revealed 1,191 contracted gene families in the *S. peregrina* , and only 568 gene families had a greater degree of expansion**(Fig. 3,** Additional file 1: Table S22**)** . The corresponding genes were identified from these gene families, including 367 contracted genes and 2,805 expanded genes, which were utilized for enrichment analyses of KEGG and GO, respectively. They mainly encoded synthesis of hydrolases, amino acid metabolism, fatty acid synthase activity, olfactory receptor activity, fibroblast growth factor receptor activity as well as chorionic membrane formation, etc. (Additional file 2: Tables S24 and S25). Besides, a total of 692 positively selected genes were identified and then analyzed by KEGG and GO enrichment analyses (Additional file 2: Table S26 and S27). These genes encoded fatty acid alpha-hydroxylase activity, epidermis morphogenesis, transferase activity, etc.

4 DISCUSSION

Compared with the most common pattern of oviparous reproduction in insects, the reproduction of ovoviviparity is relatively unique, which is defined by the ability to undergo pseudo-placental viviparous reproduction (Meier *et al.* 1999), namely, nourishing intrauterine offspring from a modified accessory gland and giving birth to larvae (Majumder *et al.* 2014). This gland is highly specialized, extending from where it connects to the uterus throughout the fat body (Attardo *et al.* 2006; Ma *et al.* 1975; Meier *et al.*1999). The glandular secretions are mainly composed of fat transferred from fat bodies during early larval development (Attardo *et al.*2006; Langley& Bursell 1980). The reproduction requires adaptive evolution of the uterus to acclimatize developing larvae, as well as adaptation of female accessory glands as nutrient synthesis and delivery system (Watanabe *et al.* 2014). In this study, genes involved in lipid metabolism are generally conserved, with gene expansions associated with fatty acid synthase, fatty acid alpha-hydroxylase activity, phospholipid metabolic process and intracellular cholesterol transport (Additional file 2: Tables S24 and S25). This pattern leads to fewer offspring per female, but a higher level of survival for the offspring (Attardo *et al.* 2006; Majumder *et al.* 2014; Meier *et al.* 1999).

Furthermore, genes involved in embryonic development have expanded significantly, mainly encoding chorionic membrane formation, fibroblast growth factor receptor activity and Dorso-ventral axis formation. Previous study implied that differentiation of ventral follicular cells is not a direct result of germline signal transduction, but relies on indirect signals from the dorsal follicle cells, providing a link between early and late events for dorsal-ventral axis formation in*Drosophila* embryos (Jordan *et al.* 2000). Despite the failure to clearly clarify the possible genetic mechanism of the reproductive pattern of *S. peregrina* , our study provides an important theoretical basis for further exploring the reproduction of ovoviviparity.

Besides, insect feeding behavior involves a broad range of activities, such as initial activation, orientation, identification and feeding (Ashworth& Wall 2010). The visual, olfactory, gustatory and neural perceptions regulate the complicated physiological processes. Olfaction plays an essential role in detecting and analyzing the semiochemicals from the environment (Field *et al.* 2000; Li& Liberles 2015). A complex and sensitive olfactory system has been developed during the long-term evolution. Necrophagous flies can colonize and breed on the decomposed corpses compared with herbivorous insects. It has been demonstrated that olfactory cues can provide a functional description of physiological mechanisms behind host choice (Carrasco *et al.*2015; Leal 2013). In this study, enrichment analysis exhibited significantly expanded genes encoding olfactory receptor activity, sensory perception of smell, sensory perception of taste (Additional file 2: Table S24).

Moreover, genes that encode neuroactive ligand-receptor interaction have expanded significantly in assembled genome. Previous study indicated that neuropeptide F (NPF) is an abundant signaling peptide in *D. melanogaster* , which play roles in feeding, reproduction, and coordinates larval behavioral changes during development (Nassel& Wegener 2011; Wu *et al.* 2003). Hence the mechanisms of host location by *S. peregrina* are of intrinsic interest, and our study sheds insight into the physiological mechanisms behind host choice.

5 CONCLUSIONS

In this study, we present high quality chromosomal-scale genome assembly of *S. peregrina* with high coverage and contiguity combining PacBio sequencing with Hi-C mapping. The publication of genomic structure and function sheds further insight into the phylogenetic diversity of flesh flies. This genome not only provides important resources for revealing the evolutionary adaptations of *S. peregrina* and other carrion-feeding species, but also it fills a gap for further study of insect evolution for the application of large-scale phylogenetic projects.

ACKNOWLEDGEMENTS

We greatly thank Prof. Lushi Chen (Guizhou Police College) for species identification of *S. peregrina* . We are also very grateful to editors and reviewers for valuable suggestions, which can greatly improve the quality of manuscript.

FUNDING

This study is supported by the National Natural Science Foundation of China (No. 81772026).

DATA ACCESSIBILITY

The data that support the findings of this study are openly available in the SRA (Sequence Read Archive) database of NCBI (National Center for Biotechnology Information) via accession numbers (SRR8416036-37, SRR9722726-27 and SRR10821909) with the Bioproject ID PRJNA509973. All versions and parameters used for software in this study are provided in Additional file 1 (Table S23).

AUTHORS' CONTRIBUTIONS

L.R., Y.G., F.M. and J.C. designed the study; L.R., L.Y., S.W. and Y.S. collected the samples; L.R., Z.B. and D.A. worked on the genome assembly, annotation and evolution analysis; L.R. and Y.S. extracted the DNA/RNA samples; L.R. wrote the manuscript; Y.G., S.C. and X.W. revised the manuscript; All authors read and approved the final version of the manuscript.

Ethics approval and consent to participate

Not applicable.

Competing interests

The authors declare that they have no competing interests.

REFERENCES

Agrawal, U. R., Bajpai, N., Tewari, R. R., & Kurahashi, H. (2010). Cytogenetics of Flesh Flies of the Genus Boettcherisca (Sarcophagidae: Diptera). *Cytologia, 75* (2), 149-155.

Alexandros, S. (2006). RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models.*Bioinformatics, 22* (21), 2688.

Allred, D. B., Cheng, A., Sarikaya, M., Baneyx, F., & Schwartz, D. T. (2008). Three-dimensional architecture of inorganic nanoarrays electrodeposited through a surface-layer protein mask. *Nano letters, 8* (5), 1434-1438.

Anstead, C. A., Korhonen, P. K., Young, N. D., Hall, R. S., Jex, A. R., Murali, S. C., . . . Gasser, R. B. (2015). Lucilia cuprina genome unlocks parasitic fly biology to underpin future interventions.*Nat Commun, 6* , 7344. doi:10.1038/ncomms8344

Ashworth, J. R., & Wall, R. (2010). Responses of the sheep blowflies Lucilia sericata and L. cuprina to odour and the development of semiochemical baits. *Medical and Veterinary Entomology, 8* (4), 303-309.

Attardo, G. M., Guz, N., Strickler-Dinglasan, P., & Aksoy, S. (2006). Molecular aspects of viviparous reproductive biology of the tsetse fly (Glossina morsitans morsitans): regulation of yolk and milk gland protein synthesis. *Journal of insect physiology, 52* (11-12), 1128-1136. doi:10.1016/j.jinsphys.2006.07.007

Bedell, J. A., Korf, I., & Gish, W. (2000). MaskerAid: a performance enhancement to RepeatMasker. *Bioinformatics, 16* (11), 1040-1041.

Belton, J. M., McCord, R. P., Gibcus, J. H., Naumova, N., Zhan, Y., & Dekker, J. (2012). Hi-C: a comprehensive technique to capture the conformation of genomes. *Methods, 58* (3), 268-276. doi:10.1016/j.ymeth.2012.05.001

Benson, G. (1999). Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Research, 27* (2), 573-580.

Birney, E., & Durbin, R. (2000). Using GeneWise in the Drosophila annotation experiment. *Genome Research, 10* (4), 547-548.

Bromberg, Y., & Rost, B. (2007). SNAP: predict effect of non-synonymous polymorphisms on function. *Nucleic Acids Research, 35* (11), 3823-3835.

Buchfink, B., Xie, C., & Huson, D. H. (2015). Fast and sensitive protein alignment using DIAMOND. *Nat Methods, 12* (1), 59-60. doi:10.1038/nmeth.3176

Burton, J. N., Adey, A., Patwardhan, R. P., Qiu, R., Kitzman, J. O., & Shendure, J. (2013). Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. *Nature Biotechnology, 31* (12), 1119.

Byrd, J. H., & Castner, J. L. (2010). Forensic Entomology: The Utility of Arthropods in Legal Investigations. 2nd Edition. *CRC Press, Boca Raton* . doi:10.4001/003.018.0221

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST+: architecture and applications. *BMC Bioinformatics, 10* (1), 421.

Carrasco, D., Larsson, M. C., & Anderson, P. (2015). Insect host plant selection in complex environments. *Current opinion in insect science, 8* , 1-7.

Castresana, J. (2000). Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Molecular Biology and Evolution, 17* (4), 540-552.

Chen, S., Zhou, Y., Chen, Y., & Gu, J. (2018). fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics, 34* (17), 884-890.

Chin, C. S., Alexander, D. H., Marks, P., Klammer, A. A., Drake, J., Heiner, C., . . . Eichler, E. E. (2013). Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nature methods, 10* (6), 563.

dos Santos, G., Schroeder, A. J., Goodman, J. L., Strelets, V. B., Crosby, M. A., Thurmond, J., . . . Gelbart, W. M. (2014). FlyBase: introduction of the Drosophila melanogaster Release 6 reference genome assembly and large-scale migration of genome annotations. *Nucleic acids research, 43* (D1), D690-D697. doi:10.1093/nar/gku1099

Dudchenko, O., Batra, S. S., Omer, A. D., Nyquist, S. K., Hoeger, M., Durand, N. C., . . . Aiden, E. L. (2017). De novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds.*Science (New York, N.Y.), 356* (6333), 92-95. doi:10.1126/science.aal3327

Dudchenko, O., Shamim, M. S., Batra, S. S., Durand, N. C., Musial, N. T., Mostofa, R., . . . Aiden, E. L. (2018). The Juicebox Assembly Tools module facilitates <em>de novo</em> assembly of mammalian genomes with chromosome-length scaffolds for under $1000. *bioRxiv* , 254797. doi:10.1101/254797

Durand, N. C., Shamim, M. S., Machol, I., Rao, S. S., Huntley, M. H., Lander, E. S., & Aiden, E. L. (2016). Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *Cell Syst, 3* (1), 95-98. doi:10.1016/j.cels.2016.07.002

Emms, D. M., & Kelly, S. (2015). OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol, 16* , 157. doi:10.1186/s13059-015-0721-2

Field, L. M., Pickett, J. A., & Wadhams, L. J. (2000). Molecular studies in insect olfaction. *Insect Mol Biol, 9* (6), 545-551. doi:10.1046/j.1365-2583.2000.00221.x

Goff, M. L., Omori, A. I., & Goodbrod, J. R. (1989). Effect of cocaine in tissues on the development rate of Boettcherisca peregrina (Diptera: Sarcophagidae). *Journal of Medical Entomology, 26* (2), 91-93.

Gongyin, Y., Kai, L., Jiaying, Z., Guanghui, Z., & Cui, H. (2007). Cuticular hydrocarbon composition in pupal exuviae for taxonomic differentiation of six necrophagous flies. *Journal of Medical Entomology, 44* (3), 450-456.

Guo, Y., Zha, L., Yan, W., Li, P., Cai, J., & Wu, L. (2014). Identification of forensically important sarcophagid flies (Diptera: Sarcophagidae) in China based on COI and period gene. *Int J Legal Med, 128* (1), 221-228. doi:10.1007/s00414-013-0923-7

Haas, B. J., Salzberg, S. L., Zhu, W., Pertea, M., Allen, J. E., Orvis, J., . . . Wortman, J. R. (2008). Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biology, 9* (1), 7.

Jordan, K. C., Clegg, N. J., Blasi, J. A., Morimoto, A. M., Sen, J., Stein, D., . . . Ruohola-Baker, H. (2000). The homeobox gene mirror links EGF signalling to embryonic dorso-ventral axis formation through notch activation. *Nat Genet, 24* (4), 429-433. doi:10.1038/74294

Jurka, J., Kapitonov, V. V., Pavlicek, A., Klonowski, P., Kohany, O., & Walichiewicz, J. (2005). Repbase Update, a database of eukaryotic repetitive elements. *Cytogenetic and genome research, 110* (1-4), 462-467.

Kalvari, I., Argasinska, J., Quinones-Olvera, N., Nawrocki, E. P., Rivas, E., Eddy, S. R., . . . Petrov, A. I. (2018). Rfam 13.0: shifting to a genome-centric resource for non-coding RNA families. *Nucleic acids research, 46* (D1), D335-d342. doi:10.1093/nar/gkx1038

Kazutaka, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability.*Molecular Biology and Evolution, 30* (4), 772-780.

Kim, J. Y., Lim, H. Y., Shin, S. E., Cha, H. K., Seo, J. H., Kim, S. K., . . . Son, G. H. (2018). Comprehensive transcriptome analysis of Sarcophaga peregrina, a forensically important fly species. *Sci Data, 5* , 180220. doi:10.1038/sdata.2018.220

Koren, S., Walenz, B. P., Berlin, K., Miller, J. R., Bergman, N. H., & Phillippy, A. M. (2017). Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Research, 27* (5), 722-736.

Korf, I. (2004). Gene finding in novel genomes. *BMC Bioinformatics, 5* , 59. doi:10.1186/1471-2105-5-59

Lagesen, K., Hallin, P., Rodland, E. A., Staerfeldt, H.-H., Rognes, T., & Ussery, D. W. (2007). RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Research, 35* (9), 3100-3108.

Langley, P. A., & Bursell, E. (1980). Role of fat body and uterine gland in milk synthesis by adult female Glossina morsitans. *Insect Biochemistry, 10* , 11-17. doi:10.1016/0020-1790(80)90033-5

Leal, W. S. (2013). Odorant reception in insects: roles of receptors, binding proteins, and degrading enzymes. *Annual Review of Entomology, 58* , 373-391.

Lee, Y. T., Chen, T. L., Lin, Y. C., Fung, C. P., & Cho, W. L. (2011). Nosocomial nasal myiasis in an intubated patient. *Journal of the Chinese Medical Association, 74* (8), 369-371. doi:10.1016/j.jcma.2011.06.001

Li, Q., & Liberles, S. D. (2015). Aversion and attraction through olfaction. *Curr Biol, 25* (3), R120-r129. doi:10.1016/j.cub.2014.11.044

Lieberman-Aiden, E., Van Berkum, N. L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., . . . Dorschner, M. O. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science, 326* (5950), 289-293.

Lowe, T. M., & Eddy, S. R. (1997). tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Research, 25* (5), 955-964.

Ma, W. C., Denlinger, D. L., Jarlfors, U., & Smith, D. S. (1975). Structural modulations in the tsetse fly milk gland during a pregnancy cycle. *Tissue Cell, 7* (2), 319-330. doi:10.1016/0040-8166(75)90008-7

Majoros, W. H., Pertea, M., & Salzberg, S. L. (2004). TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders.*Bioinformatics, 20* (16), 2878-2879.

Majumder, M., Dash, M. K., Khan, R. A., & Khan, H. R. (2012). The biology of flesh fly, Boettcherisca peregrina (Robineau-Desvoidy, 1830)(Diptera: Sarcophagidae). *Bangladesh Journal of Zoology, 40* (2), 189-196.

Majumder, M. Z. R., Dash, M. K., Khan, H. R., & Khan, R. A. (2014). The reproductive biology of flesh fly, Boettcherisca peregrina (Robineau-Desvoidy, 1830)(Diptera: Sarcophagidae). *Dhaka University Journal of Biological Sciences, 23* (1), 61-67.

Marcais, G., & Kingsford, C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of k-mers.*Bioinformatics, 27* (6), 764-770.

Martinson, E. O., Peyton, J., Kelkar, Y. D., Jennings, E. C., Benoit, J. B., Werren, J. H., & Denlinger, D. L. (2019). Genome and Ontogenetic-Based Transcriptomic Analyses of the Flesh Fly, Sarcophaga bullata. *G3 (Bethesda), 9* (5), 1313-1320. doi:10.1534/g3.119.400148

Meier, R., Kotrba, M., & Ferrar, P. (1999). Ovoviviparity and viviparity in the Diptera. *Biological Reviews, 74* (3), 199-258. doi:10.1111/j.1469-185X.1999.tb00186.x

Miura, M., Hayasaka, S., Yamada, T., Hayasaka, Y., & Kamimura, K. (2005). Ophthalmomyiasis caused by larvae of Boettcherisca peregrina.*Japanese Journal of Ophthalmology, 49* (2), 177-179.

Moriya, Y., Itoh, M., Okuda, S., Yoshizawa, A. C., & Kanehisa, M. (2007). KAAS: an automatic genome annotation and pathway reconstruction server. *Nucleic Acids Research, 35* (suppl_2), 182-185.

Nassel, D. R., & Wegener, C. (2011). A comparative review of short and long neuropeptide F signaling in invertebrates: Any similarities to vertebrate neuropeptide Y signaling? *Peptides, 32* (6), 1335-1355. doi:10.1016/j.peptides.2011.03.013

Nazni, W., Jeffery, J., Lee, H., Akmar, L. M., Chew, W., Heo, C., . . . Hisham, M. H. (2011). Nosocomial nasal myiasis in an intensive care unit. *The Malaysian journal of pathology, 33* (1), 53.

Rinkevich, F. D., Zhang, L., Hamm, R. L., Brady, S. G., Lazzaro, B. P., & Scott, J. G. (2006). Frequencies of the pyrethroid resistance alleles of Vssc1 and CYP6D1 in house flies from the eastern United States.*Insect Mol Biol, 15* (2), 157-167. doi:10.1111/j.1365-2583.2006.00620.x

Roberts, R. J., Carneiro, M. O., & Schatz, M. C. (2013). The advantages of SMRT sequencing. *Genome Biol, 14* (7), 405. doi:10.1186/gb-2013-14-6-405

Robinson, J. T., Turner, D., Durand, N. C., Thorvaldsdottir, H., Mesirov, J. P., & Aiden, E. L. (2018). Juicebox.js Provides a Cloud-Based Visualization System for Hi-C Data. *Cell Syst, 6* (2), 256-258.e251. doi:10.1016/j.cels.2018.01.001

Scott, J. G., Warren, W. C., Beukeboom, L. W., Bopp, D., Clark, A. G., Giers, S. D., . . . Liu, N. (2014). Genome of the house fly, Musca domestica L., a global vector of diseases with adaptations to a septic environment. *Genome Biol, 15* (10), 466. doi:10.1186/s13059-014-0466-3

Simao, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., & Zdobnov, E. M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics, 31* (19), 3210-3212.

Siti Aisyah, S., Baha, L., Hiromu, K., David Evans, W., & Chin, H. C. (2015). The Importance of Habitat in the Ecology of Decomposition on Rabbit Carcasses in Malaysia: Implications in Forensic Entomology.*Journal of Medical Entomology, 52* (1), 9-23.

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics (Oxford, England), 30* (9), 1312-1313. doi:10.1093/bioinformatics/btu033

Stanke, M., Steinkamp, R., Waack, S., & Morgenstern, B. (2004). AUGUSTUS: a web server for gene finding in eukaryotes. *Nucleic Acids Research, 32* (suppl_2), 309-312.

Sukontason, K., Bunchu, N., Chaiwong, T., Moophayak, K., & Sukontason, K. L. (2010). Forensically important flesh fly species in Thailand: morphology and developmental rate. *Parasitology Research, 106* (5), 1055-1064.

Thiel, T., Michalek, W., Varshney, R., & Graner, A. (2003). Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (Hordeum vulgare L.). *Theoretical and applied genetics, 106* (3), 411-422.

Tijl, D. B., Nello, C., Demuth, J. P., & Hahn, M. W. (2006). CAFE: a computational tool for the study of gene family evolution.*Bioinformatics, 22* (10), 1269-1271.

Tomberlin, J. K., Mohr, R., Benbow, M. E., Tarone, A. M., & VanLaerhoven, S. (2011). A roadmap for bridging basic and applied research in forensic entomology. *Annual Review of Entomology, 56* , 401-421.

Vurture, G. W., Sedlazeck, F. J., Nattestad, M., Underwood, C. J., Fang, H., Gurtowski, J., & Schatz, M. C. (2017). GenomeScope: fast reference-free genome profiling from short reads. *Bioinformatics, 33* (14), 2202-2204.

Walker, B. J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., . . . Young, S. K. (2014). Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement.

PloS One, 9 (11), 112963.

Walker, M. A., Pedamallu, C. S., Ojesina, A. I., Bullman, S., Sharpe, T., Whelan, C. W., & Meyerson, M. (2018). GATK PathSeq: a customizable computational tool for the discovery and identification of microbial sequences in libraries from eukaryotic hosts. *Bioinformatics (Oxford, England), 34* (24), 4287-4289. doi:10.1093/bioinformatics/bty501

Wang, Y., Ma, M. Y., Jiang, X. Y., Wang, J. F., Li, L. L., Yin, X. J., . . . Tao, L. Y. (2017). Insect succession on remains of human and animals in Shenzhen, China. *Forensic Science International, 271* (Complete), 75-86.

Wang, Y., Tang, H., Debarry, J. D., Tan, X., Li, J., Wang, X., . . . Paterson, A. H. (2012). MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic acids research, 40* (7), e49. doi:10.1093/nar/gkr1293

Wang, Y., Wang, J. F., Zhang, Y. N., Tao, L. Y., & Wang, M. (2017). Forensically Important Boettcherisca peregrina (Diptera: Sarcophagidae) in China: Development Pattern and Significance for Estimating Postmortem Interval. *Journal of Medical Entomology, 54* (6), 1491-1497.

Watanabe, Hattori, J., Berriman, M., Lehane, M., Hall, M., Solano, N., . . . Attardo, Y. (2014). Genome sequence of the tsetse fly (Glossina morsitans): vector of African trypanosomiasis. *Science, 344* (6182), 380-386.

Wells, J. D., & Stevens, J. R. (2008). Application of DNA-based methods in forensic entomology. *Annual review of entomology, 53* , 103-120. doi:10.1146/annurev.ento.52.110405.091423

Wu, G. X., Zhu, J. Y., Li, K., Gao, X., Hu, C., Cheng, J. A., & Ye, G. Y. (2013). A proteomic analysis of larval midguts of Boettcherisca peregrina in response to cadmium exposure. *Bulletin of Insectology, 66* (2), 225-229.

Wu, Q., Wen, T., Lee, G., Park, J. H., Cai, H. N., & Shen, P. (2003). Developmental control of foraging and social behavior by the Drosophila neuropeptide Y-like system. *Neuron, 39* (1), 147-161. doi:10.1016/s0896-6273(03)00396-9

Xu, L., Dong, Z., Fang, L., Luo, Y., Wei, Z., Guo, H., . . . Wang, Y. (2019). OrthoVenn2: a web server for whole-genome comparison and annotation of orthologous clusters across multiple species.*Nucleic acids research, 47* (W1), W52-W58. doi:10.1093/nar/gkz333

Xue, W., Verves, Y. G., & Du, J. (2011). A review of subtribe Boettcheriscina Verves 1990 (Diptera: Sarcophagidae), with descriptions of a new species and genus from China. *Ann. Soc Entomol. Fr., 47* , 303-329.

Yagi, M., Kosugi, S., Hirakawa, H., Ohmiya, A., Tanase, K., Harada, T., . . . Onozaki, T. (2013). Sequence analysis of the genome of carnation (Dianthus caryophyllus L.). *DNA Research, 21* (3), 231-241.

Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood.*Molecular Biology and Evolution, 24* (8), 1586-1591.

Zhuang, W., Chen, H., Yang, M., Wang, J., Pandey, M. K., Zhang, C., . . . Tang, R. (2019). The genome of cultivated peanut provides insight into legume karyotypes, polyploid evolution and crop domestication.*Nature Genetics, 51* (5), 865.

SUPPORTING INFORMATION

Additional file 1:

Figure S1 . A female adult specimen of *S. peregrina* .

Figure S2 . GC content distributions of the assembled genome.

Figure S3 . The GC depth distribution of *S. peregrina*genome.

Figure S4. Depth distribution of *S. peregrina* genome.

Figure S5 . Frequency of the *17* -mer depth distribution curve in the error-corrected reads used to the genome size estimation.

Figure S6 . Comparetive analysis of predicted annotation of gene structure characteristics between *S. peregrina* and other five dipteran species.

Figure S7 . A comparative representation of orthologous and paralogous genes aligned with other nine insect genomes.

Figure S8 . The maximum likelihood (ML) phylogenetic tree among*S. peregrina* and other nine insects.

Figure S9 . Divergence time estimation across 10 insect species.

Table S1. Evaluation for genome sequencing samples.

Table S2. Transcriptome sequencing statistics from the Illumina HiSeq to assist protein coding genes annotation.

Table S3. Genome size and heterozygosity estimation using*17* -mer analysis.

Table S4. The heterozygosity of assembled genome was estimated using SNP calling.

Table S5. Statistic of raw sequencing data based on the PacBio platform.

Table S6. Statistic of the genome assembly.

Table S7 . The completeness of the draft assembled genome was assessed using BUSCO.

Table S8. Construction of library and the Hi-C raw data from Illumina HiSeq Xten.

Table S9. Statistical analysis of Hi-C raw data after quality control.

Table S10 . Genome size of the clustered contigs in six chromosomes.

Table S11 . The completeness of the Hi-C assembled genome was assessed using BUSCO.

Table S12. Summary of protein-coding genes annotation of the genome assembly.

Table S13. Comparative statistics of the genome with other species.

Table S14. Statistic of functional annotation for predicted genes.

Table S15. Classification of repetitive sequences in the genome assembly.

Table S16. Annotation of repetitive sequences in assembled genome.

Table S17. Statistic of SSR distribution.

Table S18. Summary of SSR classification.

Table S19. Statistic of annotation for non-protein-coding genes in the genome assembly.

Table S20. Statistic of gene families among 10 insect species.

Table S21. Statistic of Single-copy orthologous genes aligned with other nine insect genomes.

Table S22. Statistic of expanded and contracted gene families among 10 insect species.

Table S23. Versions and main parameters of the software used in this study.

Additional file 2:

Table S24 . GO enrichment analysis of expanded genes.

Table S25 . KEGG enrichment analysis of expanded genes.

Table S26 . GO enrichment analysis of positively selected genes.

Table S27. KEGG enrichment analysis of positively selected genes.

FIGURE LEGENDS

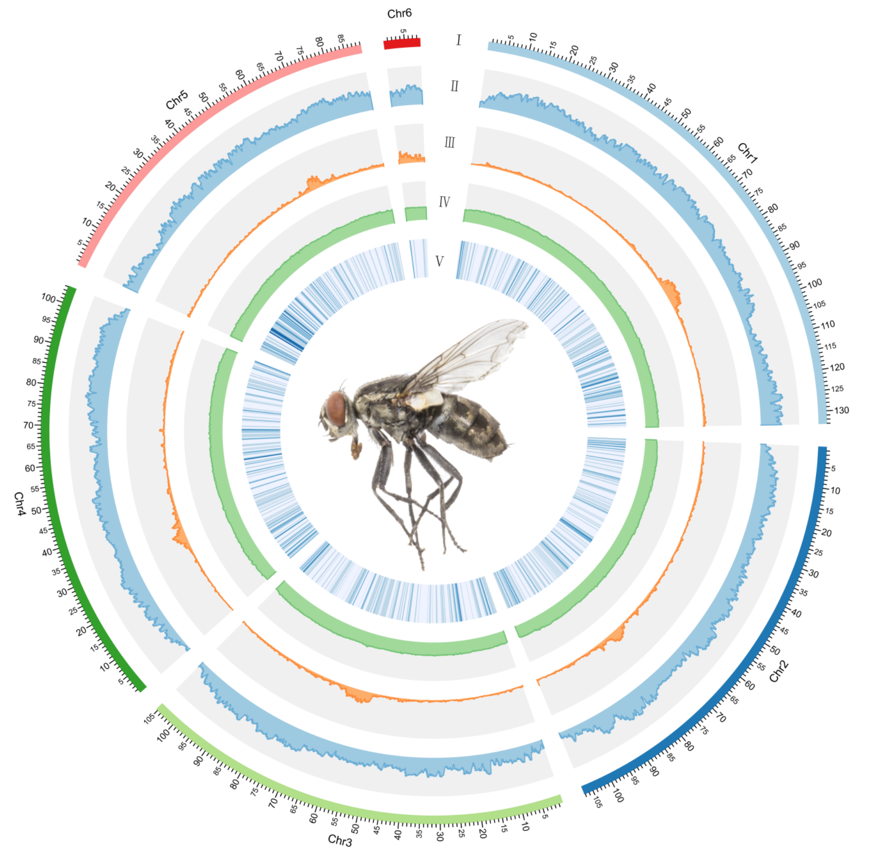
Fig. 1 Genomic landscape of *S. peregrina* . From outer to inner: (I) sizes of 6 pseudochromosomes; (II) DNA transposon content; (III) LTR transposon content; (IV) GC content (%); (V) gene density. Densities are calculated in 500 Kb windows. The photo in the circle shows an adult female.

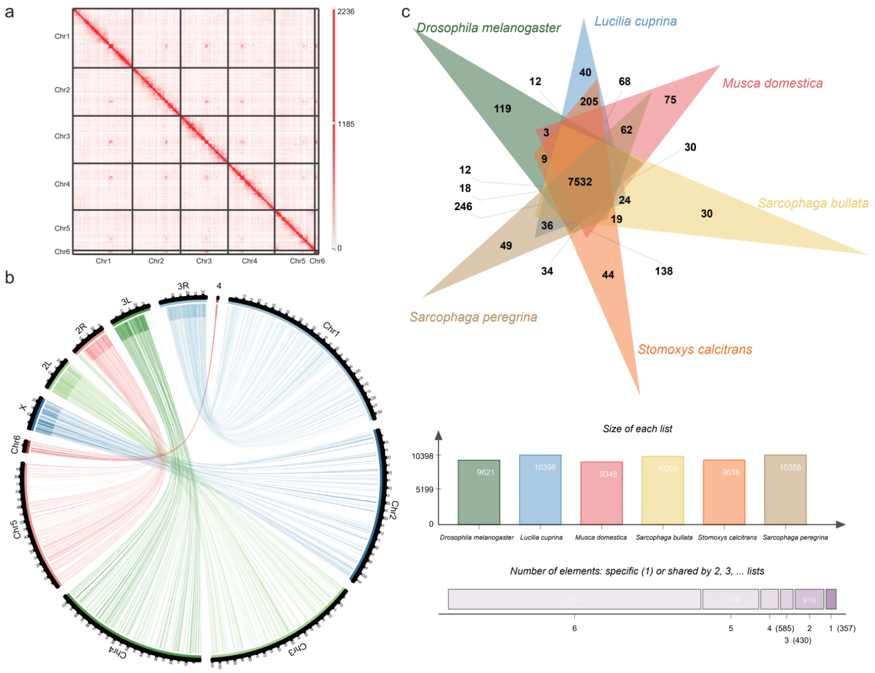
Fig. 2 Chromosome-level *de novo* genome assembly *S. peregrina* and comparative genome analysis with other species.**a** Contig contact matrix of the assembled genome. The color bar on the right shows the density of Hi-C interactions from red (high) to white (low), which are indicated number of contact links at the 100-kb resolution. **b** Chromosome collinear blocks between *S. peregrina* and *D. melanogaster* genomes. The best match across the two species is linked by lines with the same color. Chromosomes of*S. peregrina* are marked as "Chr1-6", "X, 2L, 2R, 3L, 3R, 4" respectively represents pseudochromosomes of *D. melanogaster* .**c** Venn diagram shows the distribution of orthologous clusters between *S. peregrina* and other flies (for clarity, only five species with the close evolutionary relationship to *S. peregrina*were shown). The numbers indicate gene families identified among all selected species.

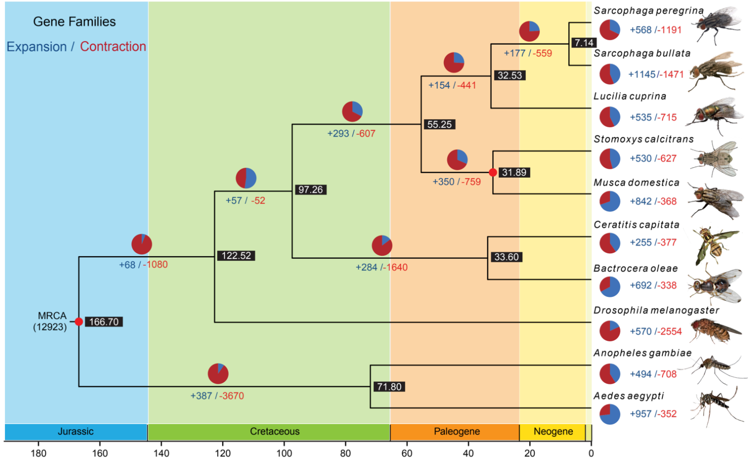
Fig. 3 Comparative genomic analyses among *S. peregrina*and nine other species. The number on the branch shows the number of expanded (blue) and contracted (red) gene families for each clade. The number near each branch indicates the number of significantly expanded (red) and contracted (blue) gene families for each clade. The black numbers show the divergence times, and two red circles indicate the calibration nodes.

Hosted file

`Revised manscript with track changes.docx` available at https://authorea.com/users/303160/articles/433277-chromosome-level-de-novo-genome-assembly-of-sarcophaga-peregrina-provides-insights-into-the-evolutionary-adaptation-of-flesh-flies


Chr6
I
II
III
IV
V
Chr1
Chr2
Chr3
Chr4
Chr5


a
Chr1
Chr2
Chr3
Chr4
Chr5
Chr6
2236
1185
0
b
3R
4
3L
2R
2L
X
Chr6
Chr5
Chr4
Chr3
Chr2
Chr1
c
Drosophila melanogaster
Lucilia cuprina
Musca domestica
Sarcophaga bullata
Sarcophaga peregrina
Stomoxys calcitrans
40
12
68
205
75
119
3
62
30
9
12
18
7532
246
24
19
30
36
49
34
44
138
Size of each list
10398
5199
0
9621
10398
9348
10200
9636
10388
Drosophila melanogaster
Lucilia cuprina
Musca domestica
Sarcophaga bullata
Stomoxys calcitrans
Sarcophaga peregrina
Number of elements: specific (1) or shared by 2, 3, ... lists
7532
1716
916
6
5
4 (585)
2
1 (357)
3 (430)



Hosted file

`Table 1.docx` available at https://authorea.com/users/303160/articles/433277-chromosome-level-de-novo-genome-assembly-of-sarcophaga-peregrina-provides-insights-into-the-evolutionary-adaptation-of-flesh-flies